2013

An original synthetic population tool applied to Belgian case: VirtualBelgium

Eric Cornelis
University of Namur, eric.cornelis@fundp.ac.be

Laurie Hollaert
University of Namur, laurie.hollaert@fundp.ac.be

Johan Barthelemy
University of Wollongong, johan@uow.edu.au

Philippe L. Toint
University of Namur

Publication Details

Cornelis, E., Hollaert, L., Barthelemy, J. & Toint, P. L. (2013). An original synthetic population tool applied to Belgian case: VirtualBelgium. Proceedings of NTTS - Conferences on New Techniques and Technologies for Statistics: The meeting place for Research in Official Statistics (pp. 695-701). Brussels: Eurostat.

An original synthetic population tool applied to Belgian case: VirtualBelgium

Abstract
A new tool for building synthetic populations, alleviating the drawbacks of classical methods, is presented. It creates both individuals and households at a quite disaggregated spatial level and simulates the temporal evolution of the built synthetic population.VirtualBelgium, this new tool, is applied on the Belgian case and is coupled with multi-agents models for simulating mobility behaviours.

Keywords
tool, original, virtualbelgium, population, case, belgian, synthetic, applied

Disciplines
Engineering | Physical Sciences and Mathematics

Publication Details
Cornelis, E., Hollaert, L., Barthelemy, J. & Toint, P. L. (2013). An original synthetic population tool applied to Belgian case: VirtualBelgium. Proceedings of NTTS - Conferences on New Techniques and Technologies for Statistics: The meeting place for Research in Official Statistics (pp. 695-701). Brussels: Eurostat.

An original synthetic population tool applied to Belgian case: VirtualBelgium

Cornelis Eric[1], Hollaert Laurie[2], Barthelemy Johan[3], Toint Philippe[4]

[1] University of Namur, naXys-GRT, e-mail: eric.cornelis@fundp.ac.be
[2] University of Namur, naXys-GRT, e-mail: laurie.hollaert@fundp.ac.be
[3] University of Namur, naXys-GRT, e-mail: johan.barthelemy@fundp.ac.be
[4] University of Namur, naXys-GRT, e-mail: philippe.toint@fundp.ac.be

Abstract

A new tool for building synthetic populations, alleviating the drawbacks of classical methods, is presented. It creates both individuals and households at a quite disaggregated spatial level and simulates the temporal evolution of the built synthetic population.VirtualBelgium, this new tool, is applied on the Belgian case and is coupled with multi-agents models for simulating mobility behaviours.

Keywords: micro-simulation, evolution, modeling

1. Introduction

The advent of micro simulation in many domains (like mobility, health, employment, etc.) increases the need for extensive disaggregate data concerning the population whose behaviour is modeled. Due to the cost of collecting this data and the existing privacy regulations, this need is often met by the creation of a synthetic population on the basis of aggregate data. While several techniques for generating such a population are known, they suffer from a number of limitations. The first is the need for a sample of the population for which fully disaggregated data must be collected, although such samples may not exist or may not be financially feasible. The second limiting assumption is that the aggregate data used must be consistent, a situation which is most unusual because this data often comes from different sources and is collected, possibly at different moments, using different protocols.

This communication highlights an original approach taking place in the stream of micro-simulation models dealing with complex realities like mobility grounded on individual level (Orcutt, 1957). Moreover we would to take into account the dynamics of population and its impacts on the studied phenomena through temporal evolutions both for individuals and households. It is a small new brick in the lineage of models starting from

DYNASIM (Orcutt et al., 1976) and in the wave of micro simulation for social sciences such as described in (Spielauer, 2009).
Our method is an innovative synthetic population generator in the class of the Synthetic Reconstruction methods, whose objective is to obviate the limitations here above mentioned.
In this paper, we will sketch in the first section why a synthetic population generator is a quite useful tool in micro-simulation. The second section will describe some current methods for generating synthetic populations. The third one will explain why such classical methods have many drawbacks for our Belgian case. Then our new generator will be described in the fourth section. Finally the fifth section will deal with dynamics of the synthetic population. In the conclusion we will highlight the use of our synthetic population tool for a prospective exercise for forecasting mobility demand and flows on an activity based model.

2. Motivation

Micro simulation means first building a data set including all the relevant characteristics of the considered agents and then simulating these agents' behaviours and updating their characteristics according to the effects of their actions. In the domain of mobility, the agents are individuals and/or households. One of the dimensions which clearly play a main role in the mobility behavior is the spatial one. Indeed mobility is a way of interacting with space, of using it. Since we plan to work at micro level, we need to deal with these interactions on a quite disaggregated spatial meshing. Otherwise variations in the agents' behaviours could not be grasped since spatial aggregation would hide any local change. That is why, for Belgium, we work at NUTS 5 (LAU2) level spreading our synthetic population amongst the 589 Belgian municipalities. Going deeper in the spatial disaggregation seems not relevant especially because it is quite unusual that data are provided at a finer level than the municipality.
Thus, we need for our micro-simulation of a population of individuals, of households characterized with a couple of attributes and located at municipality level. In an ideal world an exhaustive database would provide all the needed characteristics of the population on the given spatial meshing. We all know that such a dream never occurs. Moreover even if such a base could be available, using it would not be so easy: probably privacy rules would, in most cases, prevent using these data.
Therefore we need a methodology allowing alleviating these drawbacks. Such a method has to allow working with a picture of the population which is as close as possible to the real population. Building a synthetic (also called virtual) population is such a tool (Cirillo et al., 2012). There exist a couple of methods for such a creation of a synthetic population. We will know sketch the main ones before showing why they are not appropriate for the Belgian case, which is one of the reason for the development of our own innovative tool, VirtualBelgium.

3. Classical methods

Most of synthetic population building techniques are grounded on the IPFP (Iterative Proportion Fitting Process) principles coming from Deming and Stephan's works (Deming and Stephan, 1940) as it could be seen in (Beckman et al., 1996) (Wilson and Pownall, 1976) or (Frick and Axhausen, 2004). Roughly, it means having a representative sample of the population coupled with margins for the attributes characterizing this population. Then an individual is drawn from this sample (drawing with replacement) and placed in the category of the population corresponding to him/her, a category corresponding to the crossing of a modality for each considered characteristics. This process is repeated till the population built with the individuals so drawn exhibits margins equal to the ones known for the real population. Broadly speaking we could say that it is a building method by cloning individuals from a sample. But clearly, if some categories are not represented in the sample, they will nor be present in the built population.
This kind of "classical" method allows building a population with individuals or with households but not a "bi-level" population with individuals gathered in household. To overcome this drawback, Guo and Bhat (Guo and Bhat, 2007) proposed a new technique, once more grounded on IPFP but allowing to build both synthetic individuals and households.
Other methods are grounded on combinatory optimization (Voas and Williamson, 2001 – Huang and Williamson, 2002) but also need that a sample of the population is available.

4. Why classical methods are not suitable for the Belgian case?

If we already mentioned intrinsic drawbacks of IPFP grounded methods, mainly the absence of not observed, in the used sample, categories within the built population, we must also say that, even if these methods would be the panacea, they are not suitable for the Belgian case. Indeed, they all need a representative sample of the population and such a sample is not available for all the Belgian municipalities. Another problem is that consistent margins are necessary and our experience showed that the data which are available for Belgium are provided from different sources, were collected at different periods and therefore could exhibit inconsistencies in the margins computed from them. Examples of such incoherencies could be found in (Cornelis et al, 2005).

5. VirtualBelgium, an innovative method for building a synthetic population

All these problems and drawbacks lead us to develop an original method for building a synthetic population which not presents the problems described for the classical approaches, especially for the Belgian case. That means that this methodology needs

- to do not be fed with a representative sample of the population;
- to be suitable for case where that could exist incoherencies amongst the margins;

- to allow building a synthetic population of individuals gathered in households according to information (margins) related to one level (individuals) or the other one (households).

Since all the technical issues about this new method are fully described in (Barthelemy and Toint, 2012), we will only sketch here the main steps.

The base principle is quite simple : we generate individuals and households by randomly drawing their characteristics from the relevant distribution at the most spatially disaggregated level where they are available (e.g. some margins are available for each municipality but other ones are only provided at the district (LAU1) level). That means that the method has to deal with different levels of spatial disaggregation, making the hypothesis that, if the information are not available on the finest meshing, the distributions known for a more aggregated level could be uniformly applied for each "lower level" entities. Moreover the building method must keep the known correlation structures (i.e. comply with the cross margins which have been observed). The method also makes explicit use of both continuous and discrete optimization and used the chi^2 metric to estimate distances between estimated and generated distributions.

In concrete terms, the VirtualBelgium process goes through three steps for each of the 589 Belgian municipalities:

a. generating a pool of individuals complying with the known margins for the individuals (at municipality or district level);
b. estimating cross distributions for the households;
c. building synthetic households by drawing their members from the pool of individuals.

This new generator has been applied for constructing a synthetic population of approximately 10,000,000 individuals and 4,350,000 households localized in the 589 Belgian municipalities.

Several statistical tests as well as comparisons with IPF methods allowed us to validate our method (see (Barthelemy and Toint, 2012) for these validations).

6. Evolution of the synthetic population

The here above described method allows building a synthetic population for a given reference year. This "base" year corresponding to the period for which observations are available, for which data have been collected (however, in real cases, some extrapolation is applied since the provided margins are not always all related to the same year). For prospective exercises, for forecasting, we need to develop mechanisms allowing simulating temporal evolution of the built synthetic population. This aspect is a crucial step for dynamic micro-simulations as individuals and households change over time. For instance, individuals are aging, can get married or divorced, have children, and eventually die. All these changes can potentially have a significant influence on their behaviour. In VirtualBelgium the population can be endogenously evolved as the model is running. Since any modification of an individual agent's attribute potentially affect his/her

household, the dynamic evolution of the population is done at the household level. For every household in the simulation, the following steps are performed:

- household individuals'ages are incremented;
- possibly new babies are added to the household;
- dead individuals are removed;
- individuals'activity status and education level are updated;
- individuals are removed from the household that are wishing to leave it, and a new household is created for each of them;
- the household are splitted in two in case of divorce;
- the individuals are get married.

First, we use known fecundity and mortality rates (at municipality level) either «instantaneously » (i.e. the ones known for the reference year) or through trends (i.e. taking into account the evolution of these rates along the last years) to estimate how the pyramid of ages will evolve in the future.

For the other attributes characterizing the population (either at individuals" or households' level), different techniques could be used according to which data are available. This part of the development is currently undertaken and we think about two main methods for simulating these temporal evolutions:

- transition matrices;
- discrete choice models (RUM) (Train, 2003).

In the first case, we measure, from observations, which are the probabilities for jumping from a state to another one (e.g., for a household, the probability of jumping from the « without child » state to the « with a child » state). These probabilities could be uniform amongst the population and vary according to some other characteristics (e.g., for the here above example, the probabilities could change with the age of the family head) if data allowing such a segmentation are available. Then these probabilities are applied on the synthetic population to simulate its temporal evolution. Such a process means that we postulate that the trends in the changes of state remain constant along the years. But it is only a "business as usual" scenario which could be replaced with other prospective simulations taking into account changing trends in the future.

For the second case, the value of some attributes is estimated from the values of other ones (e.g., driving license ownership according to age, gender or diploma). This estimating relies on a discrete choice model calibrated on the available observations and then applied for the evolution (following evolution of explicative variables based on other mechanisms).

Note that the evolution time-step corresponds to one year and each sub-step corresponds to a single model.

7. Conclusion

VirtualBelgium is a platform aiming integration, consolidation and combination of multiple data sources. Indeed, using different statistical data bases and surveys, it is possible to characterize these synthetic individuals and households with relevant factors determining their behaviour in domains like mobility, employment, health, etc.

Moreover, VirtualBelgium aims at developing understanding of the evolution of the Belgian population using simulation and considers various aspects of this evolution (demographics, residential choices, activity patterns, mobility ...). To achieve this goal, VirtualBelgium uses an agent-based methodology in order to simulate the evolution of the synthetic population.
Based on statistical inputs and surveys results, it is also possible to estimate trends or to build scenarios for the temporal evolution of these factors. Therefore, it is possible through VirtualBelgium to undertake prospective exercises. For example, the mobility behaviour is modeled using an activity-based approach in which the travel demand is derived from the activities that the individuals need to perform. This approach has the advantage of reflecting the scheduling process of activities in time and space. Consequently, an agenda (or activity chain) is assigned to the individual agents in addition to their basic attributes. This agenda consist of sequence of activities. Each activity is defined by three attributes: purpose, duration, distance to be travelled to reach the activity. About 10.000 different activity chains patterns could be extracted from the Belgian national mobility survey. The duration and distance travelled for each activity are also derived from the same data source. It is assumed that every activity chains begins and end at the individual's home. Moreover the total duration of an activity chains must be less than 24 hours. These assumptions seem fairly acceptable for a large majority of the population of interest.
One can easily see that the difficulty is to merge these various models into a consistent and modular “super-model”. The modularity requirement is crucial in the sense that the agents in VirtualBelgium may receive new attributes as new data becomes available, such as the one from the Beldam (the latest Belgian mobility survey conducted in 2010). Other models must also be easily added to the simulation, for instance for employment, health, etc.
The harmonization of mobility survey as proposed by the SHANTI Cost Action would be an opportunity to build a synthetic population at European level, based on VirtualBelgium concepts, and to analyze, with a European mobility behaviours.

References

Barthelemy J., Toint Ph. (2012) Synthetic population generation without a sample, **Transportation Science**, available online

Beckman R.J., Baggerly K.A., McKay M.D. (1996) Creating synthetic baseline populations. **Transportation Research A**, 30(6), pp. 415–429

Cirillo C., Cornelis E., Toint Ph. (2012) A Model of Weekly Labor Participation for a Belgian Synthetic Population, **Networks and Spatial Economics** 12(1), pp. 59-73

Cornelis E., Legrain L., Toint Ph. (2005). Synthetic populations: a tool for estimating travel demand. In B. Jourquin, ed., **BIVEC-GIBET Transport Research Day 2005**, Vol. 1, pp. 217–235. Brussels University Press, 2005

Deming W.E., Stephan F.F. (1940) A least squares adjustment of a sampled frequency table when the expected marginal totals are known. **Annals of Mathematical Statistics**, 11, pp. 428–444

Frick M., Axhausen K. (2004), Generating synthetic populations using IPF and Monte Carlo techniques: some new results, **Technical Report¸conference paper STRC 2004**

Guo Y., Bhat C.R. (2007) Population synthesis for the microsimulating travel behavior. **Transportation Research Record: Journal of the Transportation Research Board**, 2014, pp. 92–101

Huang Z., Williamson P. (2002), A comparison of synthetic reconstruction and combinatorial optimization approaches to the creation of small-area microdata, **Working paper, Department of Geography,** University of Liverpool

Orcutt, G. H. (1957). A New Type of Socio-Economic System. **The Review of Economics and Statistics**, 39(2):pp. 116–123.

Orcutt, G., Caldwell, S., Wertheimer, R. (1976). **Policy exploration through microanalytic simulation**. Governance in Europe Series. Urban Institute.

Spielauer M. (2009), Qu'est-ce qu'une microsimulation dynamique en sciences sociales ?, Statistics Canada, Modeling Division

Train K. E. (2003) **Discrete Choice Methods with Simulation**, Cambridge University Press, 334p.

Voas D.¸Williamson P. (2001), An evaluating goodness-of-fit measures for synthetic microdata, **Geographical and Environnemental Modeling**, 5(2), pp. 177-200

Wilson A.G., Pownall C.E. (1976), A new representation of the urban system for modeling and for the study of microlevel interdependence, **Area,** 8, pp. 246-254